MR9402V1

Data Product Analysis: Early Results

White Paper
Working Paper

March 1994

Prepared Under Contract NAS5-60000

RESPONSIBLE ENGINEER

Denise Heller, Systems Engineer Date
EOSDIS Core System Project

SUBMITTED BY

Steve Fox, Office Manager Date
EOSDIS Core System Project

Hughes Applied Information Systems, Inc.
Landover, Maryland

This page intentionally left blank.

Contents

1. Introduction

2. Overview

3. Model Analysis Key Results

Figures

Tables

Appendix A. Data Pyramid Layer Definitions

Appendix B. Data Product Models, Taxonomy Phase

Appendix C. Data Type Definitions

1. Introduction

1.1 Purpose

A large number of broad decisions about the nature of the ECS Architecture are now being made, in preparation for architectural trades and ultimately for sufficient definition to support a clean design process. A significant portion of these decisions must consider information about the data to be managed by the ECS. To support those decisions, the data modeling team was tasked with detailed analysis of the products to be managed by ECS during the EOS-AM platform launch timeframe. This white paper serves to provide a summary of those results which seem most likely to have an impact on current architecture at a conceptual level, and on various portions of the architecture at a more detailed level.

1.2 Organization

This paper is organized as follows:

- Section 1 presents the purpose of the document, its organization, and logistics concerning its review, including scope and context and points of contact.
- Section 2 presents an overview of the objectives of the overall data modeling task, definition of terms pertaining to the taxonomy phase of ECS data modeling, a description of the data product analysis process that took place during this phase, and a summary of the results.
- Section 3 presents detailed results as they pertain to various architecture topics.
- Section 4 provides recommendations as to areas that may require further study in order to determine the proper path.
- Appendix B contains the partially-complete models that were prepared as the background for this analysis.

1.3 Review and Approval

This document is an informal contract deliverable approved at the Office Manager level. It does not require formal Government review or approval; however, it is submitted with the intent that review and comments will be forthcoming.

This version of the white paper will serve as input to the architecture team to support discussion in the mid-March Architecture Review/Working Group Meeting. Additional versions will be forthcoming to include comments as appropriate; another formal version release will occur just prior to the System Design Review in June. Table 1-1 shows the CDRL item(s) that will incorporate the information contained in this white paper on a more formal basis.

Table 1-1. White Paper to CDRL Migration

CDRL DID/Document Number	CDRL DID/Document Title
207/SE1	ECS System Design Specification

Questions regarding technical information contained within this Paper should be addressed to the following ECS and/or GSFC contacts:

Comments are appreciated and can be sent to

- ECS Contact

  Steve Fox
  SDPS Manager
  (301) 925-0346

- GSFC Contact

  Erich Stocker
  SDPS Technical Manager
  (301) 286-2153
  estocker@gsfcmail.nasa.gov

Questions concerning distribution or control of this document should be addressed to:

Data Management Office
The ECS Project Office
Hughes Applied Information Systems, Inc.
1616A McCormick Dr.
Landover, MD 20785

2. Overview

2.1 Data Modeling Objectives

The data modeling task is currently a three-phase effort towards determining the best organization of the data and information managed by the ECS. This activity will require revisiting, as part of evolution of the system and system requirements. In its present incarnation, its main objective is support of both the overall architecture that will be defined for presentation at the System Design Review (SDR) in June of 1994, and later support for design activities based on SDR feedback that will then be presented at the Preliminary Design Review (PDR) for the first incremental release of the system, in late fall of 1994.

In a broader sense, the data modeling activities are aimed at two key objectives:

- to provide one consistent reference for information that is deemed design-critical, so that all parties can understand the ECS data and information at a certain level of detail;
- to generate the highest (conceptual) and intermediate (logical) level models of data that can be converted into low-level physical implementation plans by those who will perform detailed design of and will ultimately implement the various release increments of the ECS software.

2.2 Definitions--Taxonomy Phase

readers: please recommend terms within this document which are not clear; they will appear in next draft.

taxonomy

model

persistence technology

data type

data pyramid

2.3 Taxonomy Phase Description

The three phases that comprise this incarnation of data modeling are a taxonomy phase, a logical modeling phase, and a physical modeling phase. The taxonomy phase, just completed, consisted mainly of the bottom-up analysis of the data and information of selected EOS instruments. These included a broad sampling from each of the TRMM, COLOR, ADEOS II and EOS-AM platforms. A variety of characteristics were researched for each of the products from this set of instruments, and recorded in "product models" for later analysis. They are included as appendix B of this white paper. By reviewing the characteristics from various angles, factors necessary to determine appropriate logical collections of data begin to appear. Based on this analysis, a taxonomy has been drawn up to represent the proposed separation for these collections. This separation strategy will be further analyzed by members of the architecture team to weigh other non-data related factors against it, and thus to refine the separation strategy.

The second phase of data modeling (logical modeling) focuses on user access requirements using various approaches. It is expected that detailed content and services needs will be revealed throughout this phase. This new information will be used to further refine, or perhaps redefine, the separation strategy. It will also be used to provide the content of nearly all levels of the data pyramid, as required by the user community. Where gaps are found between what the users need and what the product producers intend to provide, decisions will be made as to which gaps will be filled, and by whom (i.e., ECS, Producers, or Data Provider Site(s)). This task will help ensure the completeness of the architecture in capturing either the specific services and data that ECS must provide, or in describing the interfaces (via API specifications or through Delivered Algorithm Package requirements.

The instruments selected for data product analysis during this phase intended to span the variability of data types, services, and other characteristics of data to be managed by ECS. The instruments selected[1] include (listed in platform launch sequence):

- From existing Pathfinders:
  - SSM/I Polar Pathfinders
  - SSM/I Hydrology Pathfinders
  - AVHRR Pathfinders
  - TOVS Pathfinders
- From ADEOS: (1996)
  - NSCAT/SeaWinds (SWS)
- From TRMM: (1997/8)
  - CERES
  - LIS
- From EOS-AM: (1998)
  - ASTER
  - MODIS
  - MOPITT
- From EOS-COLOR: (1998)
  - EOS-COLOR/SeaWiFS II

[1]The initial selection of instruments was made within a data modeling review meeting with ESDIS personnel, and officially assigned within a memo that served as minutes of the meeting, held 21 Dec. 1993. On 10 Jan. 1994, members of the ECS Science Office, Architecture Team, PGS Toolkit development team, and data modeling team met to discuss the approach, and modifications to the list were recommended. It is the modified list that was followed.

There are, of course, several other instruments from which to choose to complete the picture; those that were chosen were felt to be sufficiently representative of a certain class of data such that leaving out another instrument that seemed also a part of that class would not significantly affect the taxonomy results. For instance, MISR was not included in this phase because its multiple band imagery seemed to present many of the same characteristics as MODIS data, while COLOR was included because of its unique purchase/non-ECS processing characteristics.

The taxonomy phase included two sub tasks in generating the product models. The first was to collect as much information as possible from documentation that we had in ECS possession, whether official documents, notes, or information held by various ECS team members familiar with the various instruments. The kind of information that was available from these sources was usually found to be of a "product description" nature.

Once this information began to be uncovered, and the modelers became more aware of what information was missing and what might be specifically applicable to this instrument, and questions for the instrument science teams were prepared. Through our Science Liaison staff at the various DAACs, we established contact with the teams in order to discover the less solid information. We found that this information was usually of a "data pyramid content" nature. It should be noted that several instrument teams met to discuss these issues, and to determine possible plans of which we were then made aware. *It is imperative that the information we found be examined with this in mind--there are very few firm plans in place for most of the products we analyzed (with the exception of the pathfinders, CERES, and LIS), simply concepts that seem appropriate at this time.*

In order to efficiently staff this phase, only four of the total pyramid layers were researched, as chosen by the architecture team representatives on that basis that these four were the most likely to vary from product to product and instrument to instrument. They were:

INVENTORY METADATA

QA STATISTICS

SUMMARY STATISTICS

BROWSE

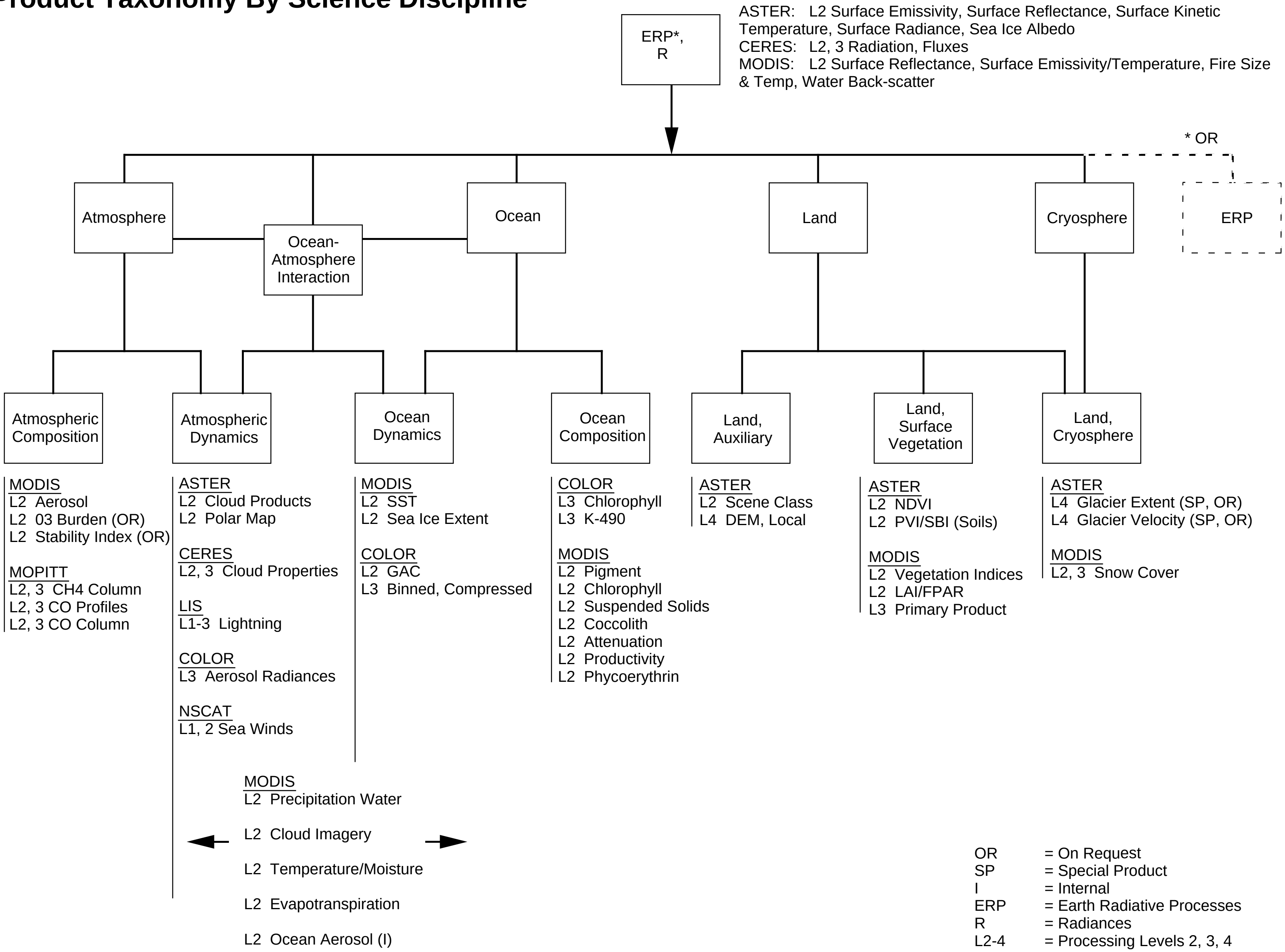
It was felt that the remaining layers, while it is important that their content be defined as part of design, do not vary enough to affect the system architecture in a significant way. Their content will thus be examined in the logical phase, driven more by user access requirements.

2.4 Summary of Results

2.4.1 Taxonomy, Logical Collections

The data product analysis process resulted in information about the product content and anticipated usage that led to the product taxonomy shown in figure 2-1. From a logical standpoint, the data products shown here would most naturally be separated in this manner.

Product Taxonomy By Science Discipline



OR	= On Request
SP	= Special Product
I	= Internal
ERP	= Earth Radiative Processes
R	= Radiances
L2-4	= Processing Levels 2, 3, 4

It is important to address in some way the common access that will likely be required to support scientific analysis of the interaction between oceans and the atmosphere. In fact, several products that measure atmospheric dynamics are either collected or generated only over the oceans, and it is assumed that this is in support of studies which will compare ocean measurements in concert with those of the spatially and temporally coincident ocean regions.

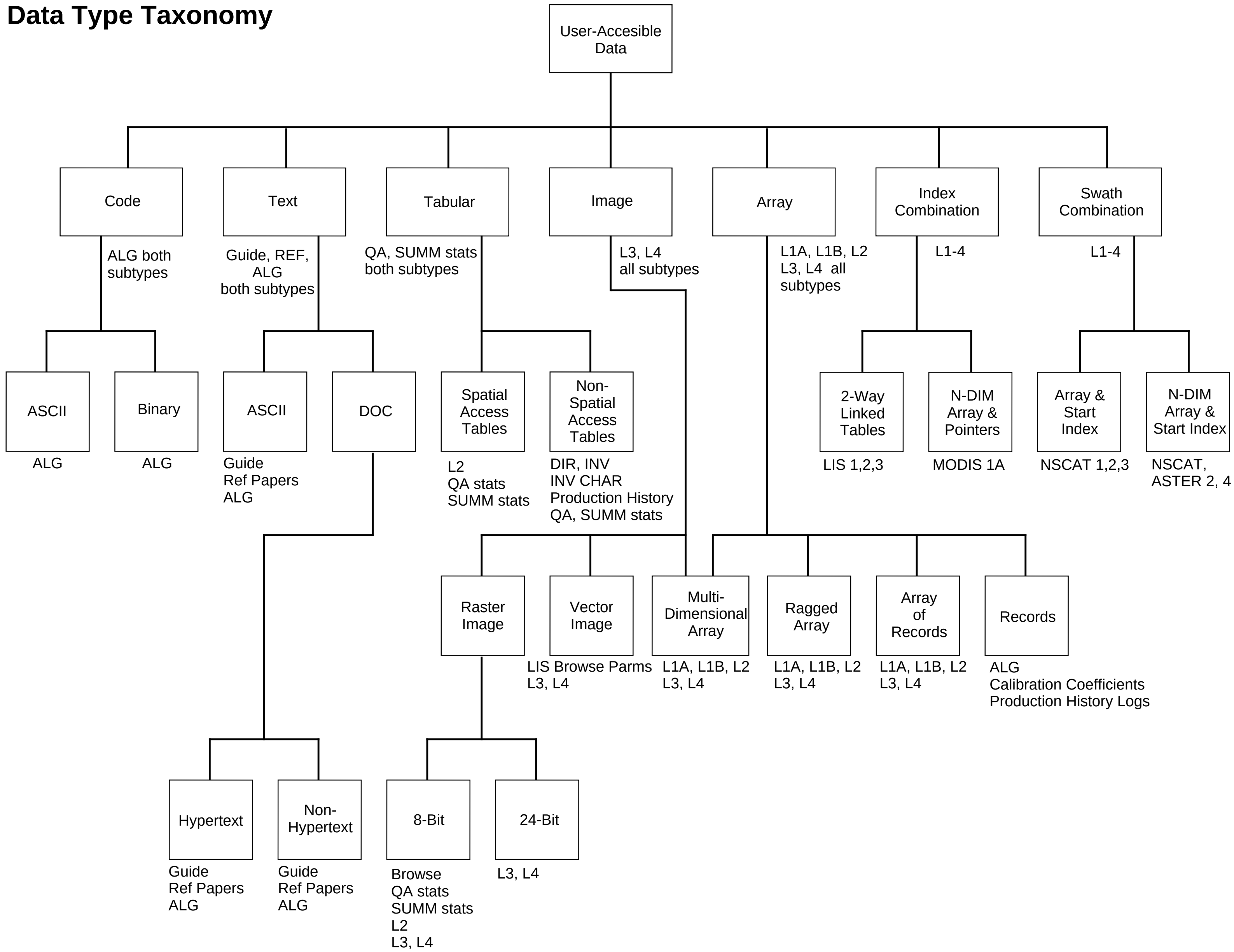
Although not well-represented in this taxonomy, there will also likely be a close relationship between products that provide Earth Radiative Processes information and almost any other earth science data. Further study should address more detailed information about these relationships; it is expected that the logical model phase, especially user scenario analysis, will provide more answers to this.

2.4.2 Taxonomy, Data Types[2]

The second taxonomy, presented in figure 2-2, shows the logical hierarchy of those data types required to support the products analyzed, as well as the layers of the data pyramid which were identified. The figure shows not only the variety of types that were found, but also how the pyramid layers map to them.

[2]Refer to appendix C for description of the data types and data type combinations that were identified in this pass. Refer also to the EOSDIS Version 0 Standard Data Format System Implementation Guidelines report dated x/93, co-authored by Ted Meyer, R. Suresh, and Karen Whalen.

Data Type Taxonomy



To look at this another way, Table 2-1 lists each pyramid layer and which data types would be required to support it. It is important to note that in several cases, product examination showed browse, QA statistics and Summary statistics existent as parameters *within* a standard product. When this is true, a particular data structure or type might not be sufficient to represent that layer; a service that extracts the data (stored within a standard product granule) to represent a particular layer must be added to cover this concept. It is expected that the logical model phase will show more completely the relationships among layers within the pyramid, and the use of services to either find or generate data for one layer from another.

Table 2-1. Pyramid/Data type Mapping, Taxonomy Phase

Pyramid Layer	Data Type Alternatives per Layer
DIRECTORY	Non-spatial access Tables
GUIDE	Text (all subtypes)
REFERENCE PAPERS	Text (all subtypes)
INVENTORY	Non-spatial access Tables
INVENTORY CHARACTERIZATION	Non-spatial access Tables
QA STATISTICS	Tabular (both subtypes), 8-bit raster image, Subset function applied to granule to extract values from index combination granule structure, Calculation function applied to tabular or subsetted values to generate statistics for graphical representation
SUMMARY STATISTICS	Tabular (both subtypes), 8-bit raster image, Subset function applied to granule to extract values from multidimensional array, Calculation function applied to tabular or subsetted values to generate statistics for graphical representation
BROWSE	8-bit raster image, bitmap (multidimensional array)
PRODUCTION HISTORY	Non-spatial access Tables, Records
ALGORITHM	Code (both subtypes), ASCII text, Records
LEVEL 4	Image (all subtypes), Array (all subtypes), Index combination (both subtypes)
LEVEL 3	Image (all subtypes), Array (all subtypes), Index combination (both subtypes)
LEVEL 2	Raster Image, Array (all subtypes), Swath Combination (both subtypes), Spatial Access Tables
LEVEL 1B	Vector Image (LIS), Array (all subtypes), Swath Combination (both subtypes), 2-Way Linked Tables (LIS)
LEVEL 1A+ANC+ENG	Array (all subtypes), Index Combination (both subtypes), Swath Combination (both subtypes)

3. Model Analysis Key Results

3.1 Logical Data Collections

EOS science support has been designed to provide a coherent view of earth science across science disciplines. The objective of the program as a whole is not only to allow but encourage interdisciplinary study in order to guide high-level policy-makers. Research results from EOS science should lead them towards decisions that might help prevent man-made *global* environmental changes, particularly global warming or ozone hole depletion.

3.1.1 Collections by discipline

There are two stages to development of the science that supports these broad policies. The first stage provides the building blocks of science that can be recombined, compared, and analyzed at the second stage. The first stage building blocks are built with the expertise and specific data related to specific disciplines and even sub disciplines within them. Examples include production of wind speed vectors, CO profiles, chlorophyll concentrations, or leaf-area indices. The use of ground truth data, higher level models, and other similar instrument products is crucial to both the creation and validation of these building blocks. This leads to a recommendation that data be logically grouped by discipline or better still by subdiscipline.

3.1.2 Collections across disciplines, flexible

The second stage is exemplified by the interaction that is inherent in the ocean and atmosphere products. As mentioned in section 2, many atmosphere products are designed such that collection and generation occurs only over ocean regions (all NSCAT products). Some MODIS atmosphere products seem specifically aimed at ocean/atmosphere interaction (e.g. Ocean Aerosols, Evapotranspiration). In addition, the sheer numbers of products and parameters that fall within either ocean or atmosphere sub disciplines suggest that this area of research is of key concern.[3] Table 3-1 shows the volumes of data which fall within the various disciplines if this taxonomy is used. Note that the greatest contributors to this weighting are the number of MODIS products as well as the number of parameters contained within CERES products in comparison to most others.

The second stage is also likely to be exemplified by interactions between the Earth Radiative Processes data and any or all of the other disciplines. As product dependencies and user access patterns are researched during phase II of data modeling, this theory will be examined more closely.

[3]This assumption is also contributed to by current literature which suggests that significant advances in understanding of global warming *and* ozone depletion can be made by understanding how the ocean acts as a sink for substantial levels of CO_2 absorbed from the atmosphere.

Table 3-1. Product/Parameter Counts Per Logical Collection[4] (1 of 2)

Discipline/Subdiscipline	Product Proc. Levels	Instrument	Product Descriptions	No. of Products	No. of Parameters
Earth Radiative Processes	L2	ASTER	Sfc Emissivity, Sfc Reflectance, Sfc Kinetic Temp, Sfc Radiance, Sea Ice Albedo	5	6
	L2,3	CERES	Radiation, Fluxes	8	161[5]
			TOTALS:	**13**	**167**
Atmospheric Dynamics	L2	ASTER	Cloud properties, Polar Cloud Map	2	14 (13 within special product)
	L2,3	CERES	Cloud properties	9 (3 AD, 6 overlap with ERP	102 (48 AD, 54 within ERP products)
	L1A-3	LIS	Lightning Data, Storms	4	30
	L2	MODIS	Precip Water, Clouds, Temp&Moisture, Evapotrans, etc	7	15
	L3	COLOR	Aerosol Radiances	1 (+1 browse)	2
	L1,2	NSCAT	Sea Winds	3	3
			TOTALS:	**27**	**166**
Ocean Dynamics	L2	MODIS	SST, Sea Ice Extent	2	2
	L2,3	COLOR	GAC Derived, Compressed	2	2
			TOTALS:	**4**	**4**
AD as OD, OC	L2	MODIS	Precip Water, Clouds, Temp&Moisture, Evapotrans, etc	7	15
			TOTALS w/ AD:	**11**	**19**

[4]The values in this table are subjective assignments that are based on this phase's data product information source, the "Summer of '93" product and parameter lists generated by the SPSO. Subsequent iterations of product lists will be evaluated for impact to the architecture prior to SDR, and will be incorporated into updates of these tables prior to PDR.

[5]This parameter total includes radiances because they are packaged within level 2 and 3 products by CERES; this may need correction given further analysis. Radiances in general (i.e. level 1B products) should be assumed to accompany the products from which they are generated, or may become a *logical* collection unto themselves.

Table 3-1. Product/Parameter Counts Per Logical Collection (2 of 2)

Discipline/Subdiscipline	Product Proc. Levels	Instrument	Product Descriptions	No. of Products	No. of Parameters
Ocean Composition	L2	MODIS	Pigment, Chlorophyll, Susp Solids, Coccolith, Attenuation, Productivity, Phycoerythrin	10	17
	L3	COLOR	Chlorophyll, K490	2 (+2 Browse)	4
			TOTALS:	**14**	**21**
Atmospheric Composition	L2	MODIS	Aerosol, O3 Burden, [,Atmos Stability]	3	4
	L2,3	MOPITT	CH4, CO Profiles, CO Column	6	6
			TOTALS:	**9**	**10**
Land, Surface Vegetation	L2	ASTER	NDVI, PVI, Soils	2	3
	L2,3	MODIS	Veg Ind, LAI/FPAR, Primary Vegetation Production	4	7
			TOTALS:	**6**	**10**
Land, Cryosphere	L4	ASTER	Glacier Extent & Velocity	2 (special products, on request only, may drop from list)	2
	L2,3	MODIS	Snow Cover	2	2
			TOTALS:	**4**	**4**
Land, General	L2,4	ASTER	Scene Classification, Local DEM	2	2
			TOTALS:	**2**	**2**

Support and encouragement of these types of interaction between disciplines calls for an additional level of logical data collections. Even within the ocean/atmosphere interaction are more logical groupings of atmospheric dynamics with ocean dynamics, or atmospheric chemistry with ocean dynamics, and so on. Earth Radiative Process data seems logically connected to atmospheric dynamics, and some products that have been assigned to this discipline easily cross into the others (e.g. land surface temperature, sea ice albedo). It is assumed that as new interactions are theorized in the future, additional couplings of the discipline collections will be required. All of these factors lead to the recommendation that the ability to access data within flexible logical "supercollections" be supported by the ECS architecture. An intermediate step towards this could be the support of four "supercollections" supporting the pairings of

- atmospheric dynamics and ocean dynamics,
- atmospheric composition and ocean dynamics,
- atmospheric dynamics and ocean composition, and
- atmospheric composition and ocean composition.

3.2 Data Types

During this phase, a small team undertook the task of assigning probable storage structures that would be required for this set of products[6]. The team used various inputs to make these decisions--including personal experience, contacts with the instrument teams, preliminary product design documents, information collected within our models, and of course engineering judgment. Additionally, a set of guidelines was determined through experience with CERES products, and through discussion with the ESDIS V0 HDF team.

3.2.1 Guidelines used to determine possible data types

The guidelines were then provided to this small sub-task team to help them with their decisions:

1. Consider the best *storage* structure. It is possible that needs involved with processing, visualization (or other access) and distribution would call for a different structure to be assigned than when just considering efficient storage. For now, we are trying to simply address storage, to establish a baseline against which User View analysis can later be compared. (This comparison will tell us, for example, that transformation to suit various usage's may be required.)
2. Images will almost always fall into one of the Raster Image subtypes or the Multidimensional Array (n-cube) type. One key to choosing between these is that Raster Images only include 8-bit and 24-bit structures; any others (such as 16 bit, for instance) would have to use a multidimensional array structure instead.
3. Note that there are 4 subtypes of Raster Images, which help to distinguish them quite readily, based on whether the image will be 8- or 24-bit, and whether it is a single band image or contains multiple bands.
4. Whenever the data description calls for variable length records (e.g., a scan that might contain a varying number of pixels per scan, with the number of pixels stated in the scanline header), a ragged array would probably be appropriate.
5. A general rule of thumb is that almost all level 1A and 1B data will best be supported by multidimensional arrays. Through process of elimination, then, determine if any of the other types might be appropriate or particularly well-suited for this product, and if none of the others seems especially so, use the multidimensional array as a default. Please be

[6]Again, see appendix C for a description of data types used in this phase.

careful not to just assume this, however! Since we are trying to identify exceptions, it is important to consider all others first, and even when left with this, consider why this might not be appropriate either. Two examples of exceptions that were uncovered by using this process of elimination are:

- The LIS data will contain logical sets of flashes, counts, and areas of coverage. Because they need to be almost "linked lists" as described by the instrument team, a unique structure turns out to be required.
- NSCAT/SeaWinds low-level data does not follow the standard across-track/along-track scan structure. Rather, data is organized within a swath sequentially along-track, but possibly in random order across track. There are relationships between the values within each swath, but they cannot be represented by assigning the orbit and time, as with most other scanners. These factors result in a mismatch with any of the existing basic structures; thus, a unique structure turns out to be required for SeaWinds as well.

6. As mentioned above, the key driver of this exercise is to uncover exceptions to the standard types, if they are required. If something doesn't fit well with the existing palette, it is better to identify it as unique, then have further analysis show that it is not, then to assume it is standard and later discover that it requires special treatment.

3.2.2 Data Types per Discipline/Collection

Figure 2-2 shows the hierarchy of data types that resulted from this task. Table 2-1 shows how each layer of the pyramid maps across instruments to these data types. What will be critical for data collection analysis in the future is how these map to the suggested collections. Table 3-2 summarizes the product to data type mapping that was determined during this phase[7]. Note that the list includes only those types applicable to the level 1A through 4 layers of the data pyramid. An assumption is being made that the full pyramid accompanying the products might vary in the types required instrument by instrument; however, insufficient data was gathered to perform that analysis at this time. Further analysis on this issue will be pursued during the logical modeling phase.

[7]The values in this table are subjective assignments that are based on this phase's data product information source, the "Summer of '93" product and parameter lists generated by the SPSO. Subsequent iterations of product lists will be evaluated for impact to the architecture prior to SDR, and will be incorporated into updates of these tables prior to PDR.

Table 3-2. Data Types Per Logical Collection, Products Only (1 of 2)

Discipline/Subdiscipline	Instrument	Data Types
Atmospheric Dynamics	ASTER	Raster Image, Multidimensional Array
	CERES	Ragged Array, Multidimensional array, Array of Records
	LIS	Array of Records, Raster Image, Multidimensional Array, Ragged Array, Index Combination
	MODIS	Multidimensional array, Image (all subtypes), ...TBD from STX
	COLOR	Multiple 2-dimensional arrays (equal angle grid)
	NSCAT	Swath combination (both subtypes), possibly multidimensional array
Atmospheric Composition	MODIS	Multidimensional array, Image (all subtypes), ...TBD from STX
	MOPITT	Ragged Array, Multidimensional Array, Index structure may be used
Ocean Dynamics	MODIS	Multidimensional array, Image (all subtypes), ...TBD from STX
	COLOR	Multiple ragged arrays with accompanying index, multiple 2-dimensional arrays
Ocean Composition	MODIS	Multidimensional array, Image (all subtypes), ...TBD from STX
	COLOR	2-D Multidimensional array, Raster Image (browse products listed as standard products)
Earth Radiative Processes	ASTER	Multidimensional array
	CERES	Ragged array, Multidimensional array

Table 3-2. Data Types Per Logical Collection, Products Only (2 of 2)

Discipline/Subdiscipline	Instrument	Data Types
Land, General	ASTER	Raster Image, Multidimensional Array
	MODIS	Multidimensional array, Image (all subtypes), ...TBD from STX
Land, Surface Vegetation	ASTER	Raster Image
	MODIS	?? TBD from STX
Land, Cryosphere	ASTER	Raster Image, Multidimensional array, ??? Need to validate with science team; these are level 4 products
	MODIS	Multidimensional array, Image (all subtypes), ...TBD from STX

Items of interest include the fact that LIS data seems to use basic structures (i.e. tables, arrays, images) but also requires some structure that helps link these disparate pieces, and assists with searching through them in two directions. For this reason, a "index combination" structure (2-way linked tables) was added to the basic taxonomy.

It is also of interest that while the MODIS scan cube at Level 1A is now easily converted to 3 multidimensional-array files, it also requires some sort of index structure for each multidimensional cube, which would contain C-language like pointers. For this reason, another "index combination" structure (n-dim array + pointers) was added to the basic taxonomy.

The Swath combinations are basically there for NSCAT's non-standard swaths: while most scanners generate data that represents collection within a swath, the NSCAT instrument does not store its samples in time-sequence. The first pixel in a scan is that which was chosen for pointing of that sweep; a location is noted for it, and a time. The remaining pixels in the scan occur randomly within the rest of the "record" that represents that scan. The location of samples is thus not calculable by knowing which orbit and the start time, as most other scanners allow. The location is rather tied to the track path, the pointing angles and so on. Further analysis will determine if this requires special indexing; for now, a special combination type (array + start index) is included in case normal array structures, or existing index structures prove insufficient.

3.3 Storage vs. Processing (Data=Services)

This concept is described in the ECS Science Data Processing Reference Architecture working paper (Ref. FB9401V1), section 4.2.5.2: "...the optimal manner for providing data to users may be dataset and product level specific, and may change over time...it may be 'cheaper' to create a level 3 data product directly from its base level 1 product each time it is requested, rather than pre-computing and storing level 2 and 3 products.". During analysis of the various levels of the data pyramid to be supported for these instruments, it became clear that the QA statistics, Summary Statistics and Browse layers could in fact be partially populated not by additional data but by services which extract this from existing standard products. For example,

- LIS level 1A includes three parameters that could be extracted and visualized as QA statistics: Errors found in the raw data (Parm 4360), Statistics of errors found in the raw data (4361) and Corrections for errors found in the raw data (4362).
- Two SSM/I Hydrology Pathfinder products contain browse products as part of the standard product: (Antenna Temperature, Unofficial product number 01, and Vegetation Index, number 07).
- LIS products also contain browse images as a parameter within the product: LIS02 parms 4369, 4370, 4371, and 4372; LIS03 parms 4377 and 4378 all qualify as browse products of their respective standard product, requiring only "subsetting" to get at them.
- CERES products contain extensive statistical content that calls for summarization to populate the summary statistics layer, by region, or for a particular cloud layer or atmospheric pressure level.

Work with both the architecture team and archive management teams during the logical modeling phase will pursue the definition of these services to a greater level of detail. Efforts can begin by looking at the above examples as a starting point.

One other aspect of storage vs. processing is the factoring in of processing dependencies. Although in-depth work in this area remains, the following list shows that there may be significant traffic between processing and archival sites if current assignments are followed:

Instrument	total products	move:	don't move
CERES	12		12
TOVS	5	3	2
SeaWiFS	13	13*	
SSM/I Polar	5		5
SSM/I Hydro	5	2	3
NSCAT	3	3	
LIS	4		4
ASTER	14(18)	(2)	14(16)
MOPITT	7		7
TOTAL:		23	49
*if SeaWiFS processed at GSFC		10	62

3.4 Pyramid Characteristic Summaries

3.4.1 Inventory Metadata

Using the CEOS Guidelines for international interoperability, several basic fields were assumed for all products, to provide search criteria[8]. These basic fields include:

Sensor/Platform

Time/Date

Location (position)

Geographical Zone

Data Quality

Orbit Characteristics

Coincident Surface Data

Instrument Attributes

Geophysical Attributes.

Where possible, product-unique requirements on the format of these items was looked into. In addition, the products and any information system that holds heritage data for those products were examined to determine product-unique descriptive fields that might be required. While substantial research is on-going regarding metadata content using user scenario analysis, existing system analysis, prototyping and a wide range of other sources, the information below is that found specifically while analyzing product-unique characteristics. This information will be folded in with the other research to recommend inventory-level metadata fields at the end of the logical modeling phase.

Table 3-3 captures variations in location and time as discovered thus far. Table 3-4 lists product- or instrument-specific fields that may be needed to adequately search for or describe the products.

[8]See "Guidelines for an International Interoperable Catalogue System", CEOS Catalogue Subgroup, Issue 2.1, April 1993.

Table 3-3. Search Criteria Variability

Instrument	Location	Time
CERES	Longitude of Spacecraft(1A)	Time is in astronomical Julian, and must be resolved with other time tags for coincidence
	Colatitude of Spacecraft(1A)	
	Regions (vary by product)	
SSM/I Polar Pathfinders	Position type plus	Time type is kept in inventory to describe content of Time Tag
	1st,2nd,3rd,4th Latitude	
	1st,2nd,3rd,4th Longitude	
	Min, Max, Middle Latitude	
	Min, Max, Middle Longitude	
	Latitude and Longitude format are used to describe fields used by granule	
SSM/I Hydrology Pathfinders	Higher level products: Lat, Lon	
	Higher level products: Grid point Lat, Grid point Lon	
	Level 1A location by start orbit, start time	
NSCAT	East,West Longitudes	
	North, South Latitudes	
LIS	Lat/Lon resolution is kept to translate	
	Attitude and ephemeris information kept	

Table 3-4. Product-Specific Search/Descriptive Fields (1 of 2)

Instrument	Fields	Instrument	Fields
CERES	product ID	**SSM/I Polar Pathfinders**	browse available
	sensor name		dataset name
	platform/source		distribution media
	campaign		first latitude
	data center		first longitude
	day/night flag		flight direction
	processing level		fourth latitude
	browse available flag		global flag
	dataset name		granule location
	[location]		IMS visible
from 1A	longitude of spacecraft/satellite		inventory name
	colatitude of spacecraft/satellite		latitude format
from FSW,SYN	grid		master directory ID
	region		max latitude
from stat products	stat time period		max longitude
	parameter		middle latitude
	parameter name be stat'd		middle longitude
			min latitude
NSCAT	granule ID		min longitude
	granule bytes		position type
	archive media type		processing level
	archive volume ID		project/campaign name
	start revolution		dataset-specific granule ID
	stop revolution		second latitude
	start time		second longitude
	stop time		spatial coverage
	dataset description		spatial resolution
	dataset east longitude		start date
	dataset west longitude		stop date
	dataset north latitude		temporal resolution
	dataset south latitude		measured parameter
	dataset long name		third latitude
	dataset name		third longitude
	dataset start time		time type
	dataset stop time		time tag
	dataset table name		type of grid
	parameter name		

Table 3-4. Product-Specific Search/Descriptive Fields (2 of 2)

Instrument	Fields	Instrument	Fields
	poly sensor data record number	**SSM/I Hydrology Pathfinders**	granule name
	processing level		storage media
	sensor name		project name
	source name		start date
	browse type id		stop date
	day/night flag		start time
	global available flag		stop time
	processing level flag		granule comments
			byte size
LIS (from TRMM)	granule ID		browse image file
	data product description		browse available
	data quality		start orbit
	file name		stop orbit
	file size		missing
	latitudinal resolution		dataset short name
	longitudinal resolution		supergranule name
	processing date	**from Rain Rate**	surface type
	processing level		lat, lon
	product name	**from Veg Ind**	Grid Point Lat
	product type ID		Grid Point Lon
	title		
	attitude information	**COLOR:SEA01**	region
	day/night flag		granule start time
	description		granule stop time
	ephemeris info		lines acquired
	number of scans in orbit		pass duration
	recording date		satellite
	scene ID		source agency
	sensor ID		filename
	spatial coverage		pathname
	start date, time		
	stop date, time		
	temporal coverage		

3.4.2 QA/Summary Statistics

Most instrument teams had little information regarding plans for quality and statistics for their products. It appears that the content and structure of this layer will be strongly driven by usage requirements of the earth science community.

3.4.2.1 QA Stats

As far as quality statistics, the SSM/I hydrology suggested that the percent of missing data per product granule could be charted for a "dataset" to help in identifying desired granules. The LIS programmable threshold for collection will be captured with each 1A granule; this threshold could be charted to show trends in the dataset for those who wish to select only those granules collected at, above or below a certain threshold. LIS level 1A data also contains three parameters which capture errors, error statistics, and corrections for the raw data. These values would likely lend themselves well to charting for the sake of selecting highest quality data. The LIS instrument team also said that the higher level products would not have quality data available, that the users would have to look at browse images of the granules to determine quality subjectively.

The MODIS classification masks are still intended for production, even though they have been removed from the standard product list (they will likely accompany the 1B product). The content of these masks are 64 bits which provide various information about the quality of the 1A data that was collected. 1-bit masks include flags for each pixel that signify: replaced dead channels vs. unreplaced noisy channels, overlapped with adjacent scan vs. no overlapped ground pixels, opaque clouds vs. transparent clouds, calculated cloud shadow vs. radiometric outlier, spatially homogeneous pixels vs. mixed pixel (mixels), land vs. water, and calculated potential glint vs. actual observed glint. 3-bit masks that contain fractional values include pixel area on ground, water fraction, opaque cloud fraction, snow/ice fraction, solar irradiance at TOA, modular transfer function significance on radiometry, and size of corrected (or uncorrected) systematic errors. These masks could either be kept or extracted into the quality statistics layer to map to some graphical representation of the 1A and 1B MODIS data.

3.4.2.2 Summary Stats

Summary statistics are somewhat more definable at this point simply because many of the products will contain statistics of some kind within the standard products. As mentioned above, CERES data contains over a hundred different parameters that represent statistics of one form or another. However, these are each captured within the products region-cell by region-cell, and parameter by parameter within those region-cells. Some products may have in the neighborhood of 40,000 cells in each granule. What is required is a means of summarizing larger groups of "regions", across all parameters within the cells, or of summarizing each parameter across all region cells, or some other roll-up of these statistics for graphical representation in this layer. This concept is shown in Figure 3-1.

LIS products at the 1A, 1B and 2 processing levels will contain statistics again as parameters within the product. For instance, LIS02 (L1B) contains the parameter 4367, Max Group Radiance, which is an array of maximum group radiance observed during each orbit at each grid location, which would map well to this layer. It also contains an array of the total observation times, an array of the number of groups observed, and a bit map of those grid boxes that had only a single event during an orbital pass. Finally, LIS03 (L2) uses a number of intermediate products to product level 3 products showing daily summaries. These include arrays of daily total flash count, daily maximum observed flash count per observation time, daily observed maximum radiance, daily observation times, daily total flash count daytime, and daily total flash count night. Also included is the daily group of global flash counts in 3 hour UTC bins. All of these statistics, if kept after processing the LIS04 (L3) product, would fit well into the summary statistics layer.

While no other instruments have yet offered this kind of detail to accompany their products, these two examples might help define a small range of what might be expected in this layer.

3.4.3 Browse and Subsetting/Subsampling, Standard and AdHoc

The approach to browse seems to vary significantly across instruments, in that browse may or may not be provided per product, or may be provided as standard products, or may be provided as parameters within some products. Most all browse products that we found discussion of were of an image nature. Non-image browse has only shown up as statistical products thus far, which may end up populating the summary statistics layer rather than serve as browse, or may serve as both. We are aware that "text" browse may need addressing, but as yet have not found a product that requires this as an analyzable sample.

Subsetting and subsampling performed by ECS must still be discussed with instrument teams specifically. Some instrument teams have stated that they do not believe that they are contractually obligated to provide browse, and thus will expect ECS to subsample as needed. They state that in past programs (Landsat was cited) browse products were obtained by users free of cost to perform science, rather than obtaining full resolution products for a fee. Explanations regarding the "free data" policy of EOS did not alleviate these fears.

CERES data will likely require subsetting/subsampling quite often because of the large numbers of parameters being packaged within granules coupled with the anticipated granule size. As mentioned above, many of the products contain ~40,000 region cells, each representing a longitude/colatitude region. In addition, parameters are kept per cloud layer (5-10), and per atmospheric pressure level (possibly up to 36 levels). It is recommended that subsetting be planned for along any of these four axes (i.e. parameter groupings, regional areas, cloud layers, and atmos levels), or even for combinations of them.

The COLOR instrument team has called for browse products as part of their standard product list. Each image will be reduced to 512 x 512 either through reduced coverage or reduced resolution. It is also likely that SEA01 will require subsetting or subsampling to extract the band values at 490μm from the resulting granules, to produce parameter 5006.

SSM/I antenna temperature granules will have 7 browse products, one for each channel, which represent horizontal polarization of 2 channels, at 720 x 720 resolution. The Sea Ice Concentration product also provides two browse images with each granule, one representing the northern hemisphere and one for the southern hemisphere.

LIS provides numerous browse images as parameters within the standard products, except for level 1A. The 1B product (LIS02) contains summary counts and events in 1.0 degree grids, global coverage. Level 2 and 3 products (LIS03 and LIS04) provide browse images for event and group counts (1° resolution total for all full resolution grid points), observations times (1˚ resolution), single events (1˚ res for all full resolution grid points), and a background image (for all frozen images, a 360x180 map of centers of the frozen images in the orbit).

3.5 Other Architecture/Design Factors

This subsection includes highlight results for various factors that may or may not ultimately affect the architecture and design. They are included to provide additional understanding of the data to be managed by the ECS.

3.5.1 Spatial Coverage

Spatial coverage must be considered both horizontally and vertically for EOS instruments. This is especially true of those instruments such as CERES and MOPITT which are aimed at atmospheric characteristics. Also, the products have varying scope of what is considered "global" coverage, for instance, global land or global ocean coverage only, global but clear sky only (MODIS) or in the case of ASTER, global coverage only where requested. ASTER also lists measurements taken at the sensor, showing "@ sensor" as the vertical resolution. (Is ASTER the only instrument that will be taking measurements at the sensor, aside from calibration values?) The variety found thus far is listed in table 3-5 below.

Table 3-5. Coverage Variability (1 of 2)

Instrument	Horizontal Coverage	Vertical Coverage
CERES	Global @	TOA
	Regional@	TOA
	Global @	Earth Locations, TOA, and Sfc
	Global @	Atmos Pressure Levels, TOA, Sfc
	Global @	Cloud Layers
TOVS Pathfinders	Global @	Surface
SeaWiFS	Oceans@	Surface
SSM/I	Global @	Surface
MODIS	Global @	Surface
	Land @	Atmos
	Ocean @	Atmos
	Global @	Clouds (level? layers?)
	Land @	Sfc
	Ocean @	Sfc
	Global @	Atmos
	Ocean Ice @	Sfc
	Global Clear Skies @	Atmos Pressure Levels

Table 3-5. Coverage Variability (2 of 2)

NSCAT	Global @	Sfc
	Ocean @	Sfc
LIS	Global @	not applicable
MOPITT	Global @	Atmos Pressure Levels
	Global @	0-15km vertical
ASTER	Global @	Sfc
	Global @	Sensor
	Land @	Sfc
	Global @	Cloud (level? layers?)
	Cryosphere @	Sfc
	Ocean Cryosphere @	Sfc

3.5.2 Grids/Projections

Several instrument teams list specific gridding schemes to be used for their products. It should be noted that of those that do, almost all use some form of equal-area grid; variability is in the resolution of grid boxes only. Information found thus far is included in Table 3-6.

No projections were named specifically for the instruments studied thus far. However, information has been obtained from the University of Colorado HRPT system development team describing various projections, their most common use, and why some would be better suited to some datasets over others. The projections include, in order of "popularity": conic, mercator, lambert conformal conic, stereographic, orthographic, transverse mercator, equid, hammer-aitoff, lambert azimuthal equal area, azimuthal equal distant, albers conic, sinusoidal, and EASE. During the logical modeling phase, the characteristics of these projections may be mapped to the products to show likely distribution of use.

Table 3-6. Grids

Instrument	Gridding Scheme
CERES	Equal Area Grids
	1.25° ISSCP-type Equal Area Grid
	1.25° Colatitude/Longitude Regions
	Equal Area Grid Boxes
LIS	Grids are uniform earth-based grid; 1.0° for browse, between 0.1° and 2.5° for standard products
MOPITT	Equal Area 22x22 km boxes
	Equal Area 22x25 km boxes
	"Standard" 1.25° colatitude/longitude regions
TOVS Pathfinders	Equal Angle Grid
SeaWiFS	Equal Area Grid
	Equal Angle Grid
SSM/I Hydrology	0.25° lat/lon Grid

3.5.3 Units

It might be important when products are obtained from multiple instruments for global change research, to either know that the parameter values are measured in different units, or be able to translate values from one to another to make them consistent. It may also be necessary to display this information as part of the results, to select the desired granules. Table 3-7 below shows the variability of units across the instruments analyzed in this phase.

Table 3-7. Units Variability (1 of 2)

ASTER	CERES	LIS	MODIS	MOPITT	NSCAT	SeaWiFS
counts	lat/lon degrees	counts	watts per square meter per sensor per channel?	watts per square meter per sensor per channel	sigma Ø in dB (decibels?)	milliwatts per square centimeter per sensor per channel
watts per square meter	watts per square meter	watts per square meter per surface region per channel?	cloud cover percent	parts per billion "v"	meters per second plus direction in degrees (vectors)	milligrams per cubic meter
elevation in meters	emissivity et al in fraction	seconds	particle phase flag: water or ice			millibars
cloud drop size in micrometers	vapor in grams per kilogram	counts per second	particle size in microns?			meters per second
water content in millimeters	drop phase flag: water or ice		millibars			ozone in DU
temperature in degrees K	cloud drop size in micrometers		fraction			
velocity in meters per second	water content in kilograms per square meter		ozone in DU (dobson units?)			
cloud height in meters	pressure in hPa		stability in C?			
"dimensionles s"--need more info on this	precipitable water in millimeters		square kilometers			
	watts per square meter per surface region per sensor?		land surface temp in degrees C			

Table 3-7. Units Variability (2 of 2)

	temperature in degrees K		cover type categorical fraction			
	water path in kilograms per square meter		fire temp in degrees K			
			milliwatts per square centimeter per sensor per channel			
			milligrams per cubic meter			
			milliwatts per square meter per sensor per channel			
			quanta per square meter per "s"			
			grams per cubic meter			
			summaries per day			
			sea surface temp in degrees K			

3.5.4 Special Considerations

Each of the instruments analyzed had certain special characteristics that may or may not be important for architecture or design, but seemed worthy of note. Most of these considerations have already been mentioned in the above sections where applicable, but are repeated here to ensure completeness.

3.5.4.1 ASTER

Four products that were previously listed on the standard products lists (AST15-18) have now been removed. They have always been considered "special products" but now may or may not be available within the EOS-AM timeframe.

Most ASTER products are on-request; this calls for "virtual metadata" as well as a means of specifying parameters that will affect processing, at the time of product "order".

ASTER is the only pointable instrument among the EOS-AM timeframe instruments, and is therefore the only instrument that will not be providing continuous temporal coverage within the dataset.

One of the standard products is a browse product (AST06), and the scene classification product (AST10) is also intended as browse.

ASTER is generating its own local DEM.

3.5.4.2 CERES

CERES products are unique in the number of parameters kept within each "product".

The product lists show several "internal" products as part of CERES standard products.

CERES products contain a variety of statistics that will lend themselves well to summarization in order to populate the summary statistics layer.

CERES is one of three instruments which are interested in vertical resolution and coverage, for atmospheric quantities/properties.

CERES provides several summary/composite products that may also populate the summary statistics layer, or may lower the volume of data transferred if users find they prefer ordering composites rather than creating their own, for trends analysis of any kind.

CERES will be using an unusual Julian date/time reference; other instruments should be consulted for their time reference preference, and translation should be investigated.

3.5.4.3 LIS

The description of products that fit the Level 1-3 classification does not map well to LIS data. The actual product structures desired combine portions of data from lower level products with those of upper level products. The actual desired structures, rather than the products listed in standard lists, must be determined.

LIS will be attempting to use many similar algorithms in an experiment that will be starting this summer. It is assumed that significant experiential information might be obtained following launch, for use in ECS PDR analysis.

Quality data, statistics, and browse images are all packaged within the LIS product descriptions, as separate parameters. If these pieces need to be stored in structures with the rest of each product, subsetting or subsampling, of a sort, can be used to "populate" the QA Stats, Summary Stats, and Browse layers.

3.5.4.4 MOPITT

This is the only mission that will focused solely on atmospheric composition.

The instrument will store 4 pixels per channel

MOPITT will operate in two modes: Stare and Burst. Burst data is one pixel over time; further investigation of this mode may show a specific example of the need for and execution of product subsampling.

3.5.4.5 NSCAT

NSCAT grids are track-based rather than earth-based. Thus, location of a "scan" cannot occur using the orbit and time. The data producers may, however, resample to map to an earth grid at level 3.

NSCAT data does not easily map to the standard processing levels. The sigma-Ø cells constitute a product that the team considers "Level 1.7", somewhere between level 1B and 2.

Presently there are plans to restrict access to NSCAT data during instrument verification.

3.5.4.6 SeaWiFS/EOS-COLOR

The EOS-COLOR instrument will utilize the ocean color algorithms that are refined by use on MODIS. Those same algorithms will be refined from those developed for SeaWiFS. Thus these three instruments have a consistency that should be found for their ocean color products, and information obtained for one should be investigated for applicability to the others.

COLOR products have not been considered as part of the ECS processing baseline. There is some rumor that ECS will in fact be responsible for processing of levels 1B and up, but currently the SeaWiFS project office is slated for all processing following a one-time purchase of 1A data for OSC.

SeaWiFS data is intended to be restricted in distribution for a five-year period. The same restriction should be investigated as to its applicability to COLOR data.

Browse products are listed as part of the standard product set for COLOR. The only product that will not have an accompanying browse product is the level 2 GAC product.

3.5.4.7 SSM/I

The hydrology pathfinders currently have supergranules associated with them. This should be investigated further.

Browse product "parameters" are contained within both the antenna temperature and vegetation index standard products.

SSM/I products have already been stored in HDF, some with as many as 15 objects of varying size and type per product, as is expected with many other products.

Coverage is limited to input data sensed over oceans only for several products (e.g. marine wind speed). The wind speed product should also be compared in resolution, quality, and data structure to the NSCAT data, which may both be best represented in swaths of some kind.

SSM/I data include one land scene classification product. Its use, format, and storage structure should be compared against scene classifications from ASTER and MODIS.

Appendix A. Data Pyramid Layer Definitions

To be supplied with later version.

Appendix B. Data Product Models, Taxonomy Phase

To be supplied with later version.

Appendix C. Data Type Definitions

The following information was developed and used by the subcontractor team, modified somewhat on the basis of experience with application to CERES, to assign candidate data types to the taxonomy phase products.

Object Code - An executable program

Data Dictionary - A data dictionary structure for defining some data content

PVL - Parameter = Value construct like ODL

Text - Text structure refers to ASCII text storage for simple documentation.

Hypertext - Text with hyperlinks

Graphics - A document with graphics and text e.g., postscript

Table - Tabular structures that would not have spatial access applied.

Basic Structures

This section will provide a conceptual understanding of the basic structures which were listed in the previous section. It is assumed that the SDF will provide explicit software support for all structures described below.

Multi-dimensional Array

Multi-dimensional arrays are n-dimensional arrays of homogenous data. Each array contains only one data type and size. All but one dimension are fixed length. This structure can be used for sensor data. Processing data can be stored in a binary table which is an instantiation of the Multi-dimensional array. The Multi-dimensional array might support the equal angle grid and sparse matrices. Examples of data types that can be stored in the Multi-dimensional array are integers of 8,16,32 ...bits, and floating point numbers of 32 or 64 bits, and possibly n bit data where n is not a multiple of 8. Figure C-1 is an example of an n-dimensional array where n= 3. The Multi-dimensional array is not limited to three dimensions. Multi-dimensional arrays may be defined with their dimensions in any order to optimize the storage for a certain method of access or to emulate any style of interleaving (BSQ, BIP, BIL)

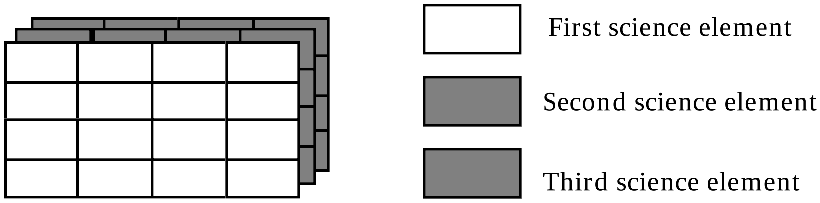


Figure C-1 An example of a Multi-dimensional Array

Image

An image is a two dimensional array of spatially organized measurements. Images typically contain 8- or 24-bit pixels. Image data may contain bands in different spectral wavelengths. An 8-bit image is generally associated with a palette. Figure C-2 is an example of an image structure.

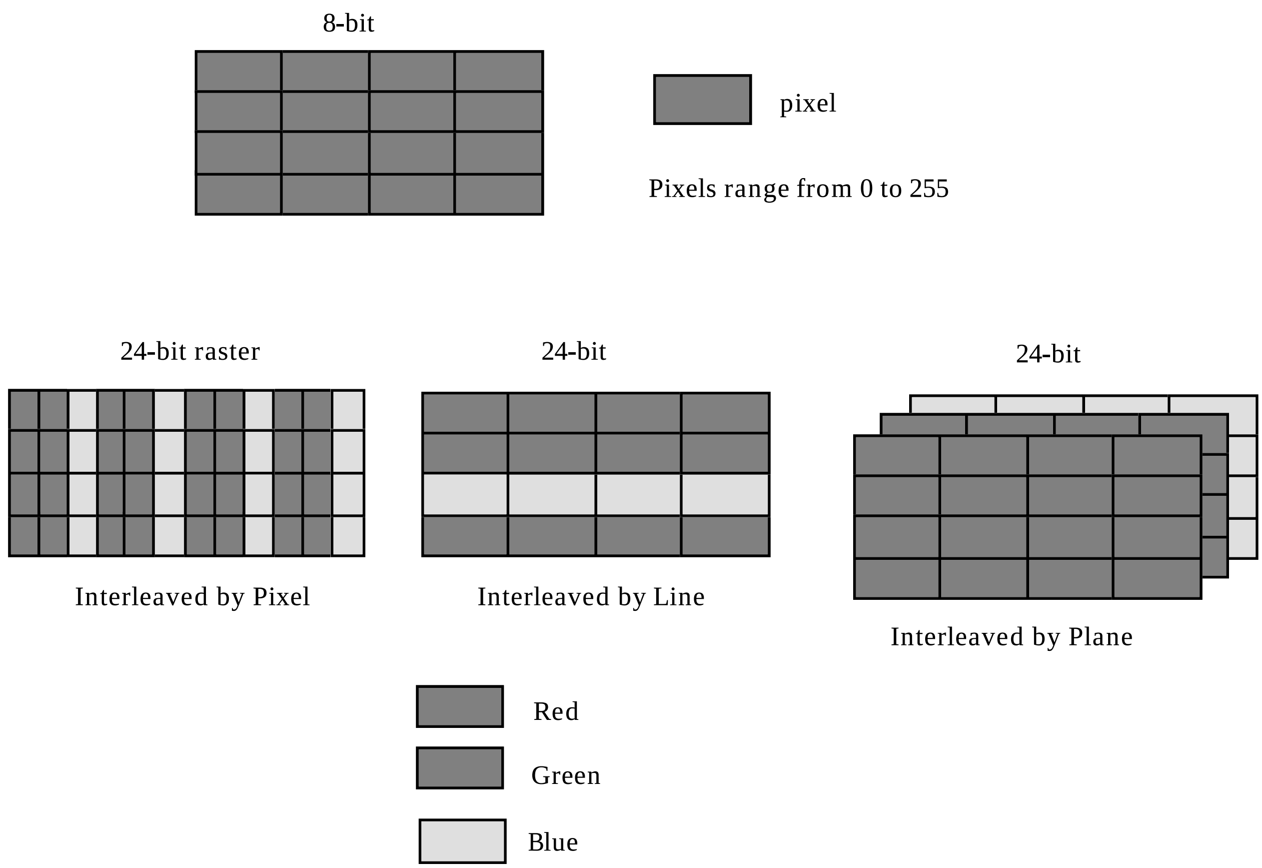


Figure C-2 An example of an image

Palette

A palette consists of an 8 bit lookup table which associates a color with each of 256 possible pixel values which can be stored in an 8 bit image.

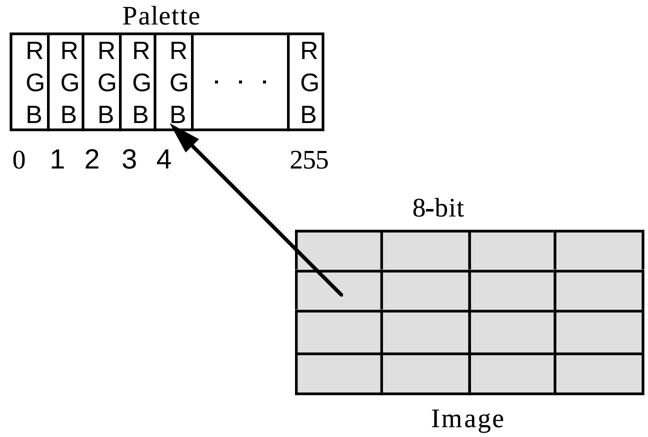


Figure C-3 An example of a palette

Ragged Array

A ragged array is a multidimensional array for storage of homogenous binary data with variable length along one direction. A row may contain multiple science elements of the same data type and size. This structure supports the equal area grid. Examples of data types that can be stored in the ragged array are integers of 8,16,32 ...bits, and floating point numbers of 32 or 64 bits,, and possibly n-bit data where n is not multiple of 8. Figure C-4 shows an example of a 3 dimensional ragged array with the variable length dimension shown horizontally.

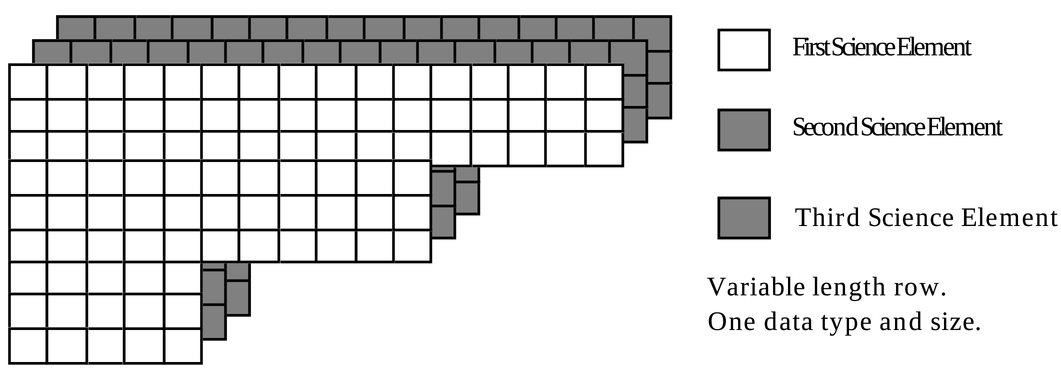


Figure C-4 An example of a ragged array

Data may be interleaved by plane where each plane consists of a different science element (above). Data may consist of the same science element for all planes in the ragged array. Data may be interleaved by science element with multiple science elements per row.

Array of Records (Table)

An array of records is a multi-dimensional array for storage of heterogeneous binary data. An array of structures may contain character, integer and floating point data. This structure may support point data. Table is an instantiation of array of records.

A table is a one-dimensional instantiation of the array of records, in which a row defines a heterogeneous structure. Each column can be of any allowable data type. Example: spreadsheets.

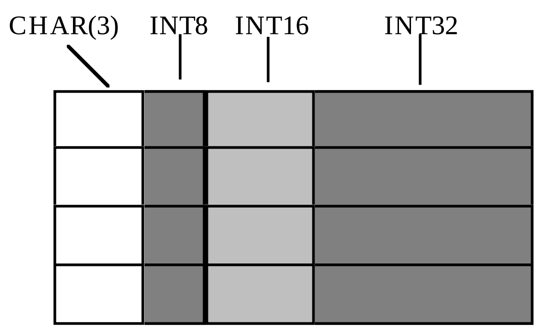


Figure C-5 A table as an array of records

Index Structure

An index structure consists of a table for indexing location and other information pertaining to the science data. This structure may be used to support point data.

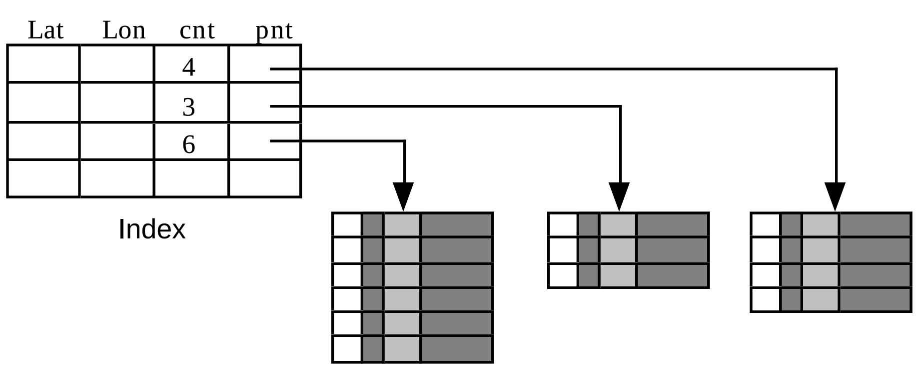


Figure C-6 An example of index structure

Vector Data

Structures mostly include Each data contains location. The set of data can be a point, node, line, arc, polygon.

Unique Structures

MODIS level 1A data structure is an example of complex structures. It has 10 detectors, 83 bands of data in addition to geolocation, calibration and radiometric data.

LIS

Swath Data

Swath data is best described by examining the scenario in which it arises. Swath data is most often produced by an orbiting scanning sensor which has a set of detectors scanning in the cross-track direction. The motion of the satellite (by definition, in the along-track direction) causes the footprint of the data to form a "ribbon" centered on the sub-nadir track. In the case of polar orbiting satellites, this ribbon will continually wrap around the Earth from pole-to-pole.

The swath data structure is probably most applicable to products at levels 2 and 3.